به نام خدا

تمرین ۲، گزارش مربوط به دیتاست ۲

علیرضا نجاتی، ۹۸۲۲۲۱۰۴

مقدمه:

این دیتاست مربوط به اطلاعات آگهی های اجاره خونه های مختلف در مناطق مختلف کشور آلمان می باشد. در این دیتاست ویژگی های یک خونه و برخی از ویژگی های مربوط به آگهی آورده شده است. در این تمرین می خواهیم با استفاده از سه ویژگی خواسته شده ('heatingType' , 'serviceCharge' , 'telekomUploadSpeed') مدل های مختلف را بسازیم و نتیجه ها را با یکدیگر مقایسه کنیم. همچنین در این تمرین مدل Linear Regression را با تابع خطا MSE از پایه پیاده سازی می کنیم.

توضیحات:

۶ سلول ابتدایی مربوط به اتصال به کگل، دریافت دیتاست، دسترسی کتابخانه های مورد نیاز و ساخت دیتافریم از روی دیتاست می باشد.

۱- پیش پردازش داده ها :

در این قسمت پیش پردازش و پاکسازی داده ها در آن انجام شده است که مشابه پیش پردازش انجام شده در تمرین اول است. ابتدا تعداد داده های null را برای هر ستون مشخص کرده و سپس آن ستون هایی که بیشتر از ۵۰ درصد داده های آن ها null هستند را حذف میکنیم. سپس با توجه به اطلاعاتی که درون دیتاست قرار داشت، آن ستون هایی که اطلاعات مفیدی نداشتند و عملا قیمت اجاره خانه به آن ها ربطی ندارد را حذف کردیم. در قسمت بعدی داده های عددی null را با میانگین ستون مربوطه پر می کنیم. سپس به سراغ داده های null در ستون های categorical می رویم و آن ها را با داده با بیشترین تکرار در آن ستون جایگزین می کنیم. دیگر داده پوچی در دیتاست وجود ندارد. در ادامه داده هایی که احتمالا به اشتباه در دیتاست وارد شده اند را پیدا می کنیم. مثلا داده هایی که قیمت اجاره آن ها برابر با ۰ است و داده هایی که مساحت آن برابر با ۰ است را پیدا کرده و سطر مربوط به آن ها را حذف می کنیم. داده های پرت عددی را به وسیله روش z_score پیدا می کنیم و آن ها را از دیتاست حذف می کنیم. ستون های catrgorical که مقادیر مختلف زیادی را می گیرند را پیدا و آن ها را حذف می کنیم. تعداد داده های دارای مشابهت را پیدا کرده سطر مربوط به آن ها را حذف می کنیم.

۲- تسک اول :

در این قسمت ابتدا فیچر های خواسته شده به همراه فیچر هدف یعنی totalRent را از دیتافریم اصلی جدا کرده و دیتافریم جدید را می سازیم. سپس با استفاده one hot encoding متغیر های categorical را به متغیر های عددی تبدیل می کنیم. سپس داده های آموزش و تست را به نسبت .۰۸ به ۰,۲ جدا می کنیم و در مرحله ی بعد آن ها را با روش MinMax هم مقیاس یا به اصطلاح نرمال می کنیم. سپس کلاس مدل Linear Regression خودمان را با تابع های batching برای تکه تکه کردن داده ها، تابع loss برای بدست آوردن مقدار loss، تابع predict برای پیش بینی مدل از روی داده های تست، تابع mse برای بدست آوردن مقدار Mean Squared Error و تابع train برای fit کردن مدل بر روی داده های آموزش پیاده سازی می کنیم. مدل را بر روی داده ها اجرا می کنیم و نتیجه را گزارش می کنیم. مقدار MSE برای این حالت برابر با ۲۱۲۸۶۴ بدست آمد.

۳- تسک دوم :

در این قسمت مدل Linear Regression با استفاده از پکیج sklearn پیاده سازی می کنیم و نتیجه را چاپ می کنیم. مقدار MSE برای این حالت برابر با ۲۱۲۶۶۰ بدست آمد که اختلاف ۲۰۰ واحدی با این مقدار برای مدل پیاده سازی شده توسط خودمان دارد.

۴- تسک سوم :

در این بخش می خواهیم مدل های Lasso Linear Regression و Ridge Linear Regression را با استفاده از پکیج های sklearn پیاده سازی کنیم. برای مدل Lasso مقدار MSE برابر با ۲۱۲۶۷۸ بدست آمد که مقداری بیشتر از مدل اصلی با پکیج است. برای مدل Ridge مقدار MSE برابر با ۲۱۲۶۶۰ بدست آمد که همان مقدار MSE مدل اصلی با پکیج است.